# Maximum likelihood estimering med logistisk regression som eksempel

Lolle, Henrik Lauridsen

*Publication date:*
2020

# Maximum Likelihood estimering med logistisk regression som eksempel

Henrik Lauridsen Lolle

*Sidst revideret: januar 2020*

I det følgende vil jeg kort, og på rent praktisk facon, vise logikken i *maximum likelihood* estimering eksemplificeret ved binær logistisk regression i statistikprogrammet Stata.

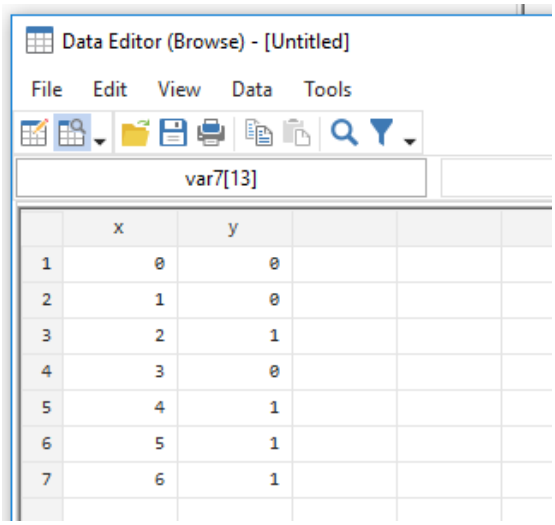
Til brug i eksemplet har jeg til en start genereret et lille datasæt med blot syv cases. Heri er der en intervalskaleret uafhængig variabel, der antager værdier mellem 0 og 6, samt en afhængig dummy-variabel. Jeg har ordnet analyseenhederne efter værdierne på den uafhængige variabel, sådan at det umiddelbart fremgår ret tydeligt, at der er en positiv sammenhæng, hvor jo højere *x*-værdi, des større sandsynlighed for succes ($y = 1$).

```
. input x y

            x           y
  1. 0 0
  2. 1 0
  3. 2 1
  4. 3 0
  5. 4 1
  6. 5 1
  7. 6 1
  8. end

.
end of do-file
```

Programmet giver følgende datasæt:




| | x | y |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 1 | 0 |
| 3 | 2 | 1 |
| 4 | 3 | 0 |
| 5 | 4 | 1 |
| 6 | 5 | 1 |
| 7 | 6 | 1 |

Jeg foretager nu en logistisk regression med *y* som afhængig variabel og *x* som uafhængig variabel:

```
. logit y x

Iteration 0:   log likelihood = -4.7803567
Iteration 1:   log likelihood =  -2.558931
Iteration 2:   log likelihood = -2.4934174
Iteration 3:   log likelihood = -2.4911632
Iteration 4:   log likelihood = -2.4911595
Iteration 5:   log likelihood = -2.4911595

Logistic regression                             Number of obs     =          7
                                                LR chi2(1)        =       4.58
                                                Prob > chi2       =     0.0324
Log likelihood = -2.4911595                     Pseudo R2         =     0.4789

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    1.250679   .8834025     1.42   0.157    -.4807582    2.982116
       _cons |   -3.110739   2.517373    -1.24   0.217      -8.0447    1.823221
------------------------------------------------------------------------------
```

Variablen *x* har ikke nogen statistisk signifikant effekt ifølge den såkaldte Wald-test (p=0,157). Derimod er Likelihood-ratio-testen for modellen, som jo i dette tilfælde er det samme som en test af variablen x, statistisk signifikant på 0,05-niveau (p=0,032). De to typer af test vil normalt ligge tæt, men LR-testen har i små datasæt mere *power* end Wald testen. Med så få cases ville jeg dog normalt ikke diskutere statistisk signifikans. Jeg ville fx være lidt utryg over, om de syv cases nu også var helt tilfældigt udvalgt.

Fokus i denne tekst er imidlertid ikke på den statistiske signifikans i eksemplet, og ej heller på forskellen mellem Wald-testen og LR-testen. Fokus er derimod på en forståelse af, hvad likelihood-værdier er for en størrelse, samt på, hvad det vil sige at maksimere likelihood-værdien.

En likelihood-værdi er i sammenhæng med statistisk analyse en bestemt form for sandsynlighed, nemlig den modelestimerede sandsynlighed for at få stikprøvens værdier på den afhængige variabel, givet værdierne på den uafhængige variabel. I *maximum likelihood* drejer det sig om at maksimere likelihood-værdien, dvs. konstruere den model, der passer bedst med data i stikprøven. Det er en proces, der iterativt gennemløber en estimering af likelihood-værdi, hvor der efter hvert resultat ændres en smule på modellens koefficienter, indtil at der nås et toppunkt (maximum likelihood).

I output fra statistikprogrammer rapporteres der som oftest den naturlige logaritme til likelihood-værdien, også kaldet Log likelihood. Denne fremgår i Stata output ovenfor lige over tabellen samt ud for iteration 5, og den er her lig med -2,4911596. Selve likelihood-værdien (sandsynligheden) er derfor lig med:

$$exp(-2{,}4911595) = 0{,}0828139$$

Hvis modellen, der fremgår i tabellen under "Coef.", er den, der gælder i populationen, vil der være en sandsynlighed på cirka 0,08 for at få lige netop denne stikprøves *y*-værdier, givet værdierne på *x*. Her ved denne femte iteration (og altså med den model, der er præsenteret resultater fra) kan Stata ikke finde nogen *anden* model, der passer bedre på data i stikprøven. Men hvordan beregnes så denne likelihood-værdi for hele sættet af *y*-værdier, givet *x*-værdierne? Det gør man ved at tage udgangspunkt i modellens ligning:

$$ln\left(\frac{P[y=1]}{1-P[y=1]}\right) = -3{,}110739 + 1{,}250679x$$

Med ovenstående formel kan jeg estimere sandsynlighed for succes, dvs. $P[y = 1]$, for samtlige syv analyseenheder. Og jeg kan derfor også nemt estimere sandsynligheden for, at det *ikke* er succes, dvs. $P[y = 0]$. Sandsynligheden for ikke-succes er blot én minus sandsynligheden for succes, fordi den samlede sandsynlighed (for enten 0 eller 1) er lig med 1. Hvis man vil estimere den samlede likelihood-værdi for *hele* stikprøven, skal man starte med, for hver analyseenhed, at estimere likelihood (sandsynlighed) for at få den empirisk fundne *y*-værdi, givet modellen og givet værdi på *x*. Det har jeg gjort manuelt i tabellen nedenfor[1].

| A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|
| Case-nr. | x-værdi | Logitværdi -3,110739+1,250679(x) | Odds =exp(Logit) | P(y=1) =Odds/(1+Odds) | P(y=0) =1-P(y=1) | y-værdi | Likelihood for y-værdi |
| 1 | 0 | -3,1107 | 0,0446 | 0,0427 | 0,9573 | 0 | 0,9578 |
| 2 | 1 | -1,8601 | 0,1557 | 0,1347 | 0,8653 | 0 | 0,8653 |
| 3 | 2 | -0,6094 | 0,5437 | 0,3522 | 0,6478 | 1 | 0,3522 |
| 4 | 3 | 0,6413 | 1,8989 | 0,6550 | 0,3450 | 0 | 0,3450 |
| 5 | 4 | 1,8920 | 6,6325 | 0,8690 | 0,1310 | 1 | 0,8690 |
| 6 | 5 | 3,1427 | 23,1653 | 0,9586 | 0,0414 | 1 | 0,9586 |
| 7 | 6 | 4,3933 | 80,9098 | 0,9878 | 0,0122 | 1 | 0,9878 |
| | | | | Samlet likelihood-vædi (multiplikation af de enkelte likelihood-værdier): | | | 0,0829 |

Jeg vil tage et par cases fra tabellen som eksempler. I case nummer 3 i tabellen er $x = 2$, og på baggrund af dén *x*-værdi samt modellens ligning bliver der i kolonne C beregnet en logitværdi på -0,6094[2]. Odds for succes er den eksponentielle værdi heraf[3], og den er i kolonne D beregnet til 0,5437. Sandsynligheden for succes ($y = 1$) kan beregnes som oddsene divideret med 1 + oddsene, som i kolonne E beregnes til 0,3522. Med andre ord: ud fra modellen vil der ved en *x*-værdi på 2 være en sandsynlighed på cirka 35 pct. for at y er lig med 1. Af kolonne G fremgår, at *y* faktisk *er* lig med 1 for case nummer 3. Sandsynligheden for at få den i stikprøven fundne *y*-værdi for case nummer 3 er altså lig med 0,3522, hvilket også er noteret i kolonne H som likelihood-værdi.

Hvis vi i stedet for ser på case nummer 2, hvor *x* er lig med 1, så er sandsynligheden for succes, dvs. $y$=1, beregnet til 0,1347 i kolonne E. Imidlertid er den i stikprøven fundne *y*-værdi her lig med 0, så vi skal i stedet for bruge sandsynligheden for $y = 0$ for at komme frem til likelihood i kolonne H. Sandsynligheden for $y = 0$, når *x* er lig med 1, er 1 minus sandsynligheden for succes, dvs. 1-0,1347 = 0,8653. Derfor noteres netop det tal som likelihood i kolonne H.

Nu er det imidlertid ikke likelihood for den enkelte case, jeg er interesseret i, men derimod den samlede likelihood for stikprøven. Som jeg skrev indledningsvis, angår likelihood-værdien, der angives i Stata-resultaterne, den modelberegnede sandsynlighed for at få stikprøvens værdier på den afhængige variabel, givet værdierne på den uafhængige variabel. Dvs. at det drejer sig om sandsynligheden for det første udfald på y, det andet udfald på y, det tredje udfald på y osv. *til sammen*. Udfaldene er uafhængige af hinanden, og derfor er den samlede sandsynlighed lig med multiplikationen af samtlige enkeltudfald[4]. I den nederste række i tabellen er den samlede

[1] Det er også muligt i Stata ved hjælp af postestimeringskommandoen `predict` at gemme de modelestimerede sandsynligheder for succes for hver case som en variabel.

[2] $-3{,}110739 + 1{,}250679(2) = -0{,}6094$

[3] Med det naturlige tal *e* som base.

[4] Se fx Alan Agresti (2018). *Statistical Methods for the Social Sciences*. Pearson, 5. udgave p. 81; eller: https://www.wyzant.com/resources/lessons/math/statistics_and_probability/probability/further_concepts_in_probability

likelihood-værdi beregnet ved netop at multiplicere de enkelte likelihood-værdier. Det giver samme resultat som den likelihood-værdi, der ovenfor blev beregnet for modellen på baggrund af den af Stata rapporterede log likelihood-værdi (på nær i sidste decimal, som skyldes manglende præcision i de manuelle udregninger, hvor der er foretaget afrundinger).

Det er i øvrigt også værd at bemærke, at en stikprøve ikke skal indeholde synderligt mange cases, førend den samlede likelihood-værdi bliver ekstremt lille, da der undervejs i estimeringen heraf bliver multipliceret med et tal mindre en 1 for hver enkelt case (sådan som jeg gjorde ned gennem kolonne H). Kloner man fx datasættet i eksemplet og lægger klonen til, så det indeholder de samme cases to gange, dvs. 14 cases i alt, vil koefficienterne blive de samme, og likelihoodværdierne for de enkelte cases vil også være de samme. De vil blot blive gentaget to gange, og den samlede likelihoodværdi vil være 0,0829×0,0829 = 0,0069 (hvis man tager udgangspunkt i tabellen ovenfor med de små afrundingsfejl).

Som det fremgår af resultaterne af logit-kommandoen, der blev præsenteret ovenfor, blev der i søgningen efter *maximum* likelihood foretaget fem iterationer efter start-estimationen (iteration 0). Det fremgår, at værdien for log likelihood hele vejen op til iteration 5 kommer tættere på 0, hvilket også vil sige, at likelihood-værdien op gennem iterationerne *stiger*. I Stata kan man fremkalde modellerne fra de tidligere iterationer ved at benytte den option, der hedder trace:

```
. logit y x, trace
```

*Udsnit af resultater:*

```
Iteration 2:
Parameter vector:
            y:          y:
            x       _cons
r1   1.202774  -2.945665

                                                log likelihood = -2.4934174
```

Hvis man fx prøver at benytte ligningen til iteration 2 i beregningerne af samlet likelihood-værdi, vil man ende med en værdi på exp(-2,4934174)=0,0826, altså en anelse mindre end *maximum* likelihood for disse data. Med ligningerne til iteration 3 og 4 vil man komme endnu tættere på maximum likelihood; så tæt på at der med fire decimaler efter kommaet ikke vil være forskel. Endelig i iteration 5 kan der med de meget små skridt, som Stata tager, ikke ændres noget i ligningen, uden at likelihood-værdien mindskes; deraf betegnelsen *maximum likelihood*.

Den første beregning af log likelihood, der foretages i logistisk regression, vil pr. default være for en model, hvor effekten fra den uafhængige variabel (eller *de* uafhængige variabler) er lig med 0, dvs. hvor der reelt kun er et konstantled i ligningen. Log likelihood for iteration 0 (eller den såkaldte nul-model) er i eksemplet lig med -4,7802567, dvs. en likelihood-værdi på exp(-4,7802567)=0,0084, altså betydeligt mindre end maximum likelihood på 0,0826. Likelihood-værdierne kan også benyttes som statistisk signifikanstest. Jeg vil dog ikke i denne tekst komme nærmere ind på det, da eksemplets datasæt under alle omstændigheder vurderes for lille til at kunne foretage såkaldte *likelihood ratio test*. Se derfor herom i en senere note.